# SCIENCE AND TECHNOLOGY TEXT MINING: CITATION MINING OF DYNAMIC GRANULAR SYSTEMS

**By**

**Dr. Ronald N. Kostoff, Office of Naval Research**
**Arlington, VA, USA**

**Dr. J. Antonio del Río, Centro de Investigación en Energía, UNAM**
**Temixco, Mor. México**

**Lic. Esther Ofilia García**
**Centro de Investigación en Energía, UNAM**
**Temixco, Mor. México**

**Lic. Ana María Ramírez**
**Centro de Investigación en Energía, UNAM**
**Temixco, Mor. México**

**Mr. James A. Humenik, NOESIS, Inc.**
**Rockville, MD, USA**

*(The views in this paper are solely those of the authors, and do not necessarily represent the views of the U.S. Department of the Navy or any of its components, the Universidad Nacional Autonoma de Mexico, or Noesis, Inc.)*

**KEYWORDS**: citation mining; text mining; data mining; bibliometrics; scientometrics; computational linguistics; citation analysis; research impact; research evaluation; clustering; phrase frequency

## I. ABSTRACT

**Background:** Research sponsors, evaluators, managers, and performers have strong motivations in insuring that their research products reach the intended audience. Further, it is important to understand the infrastructure characteristics of the specific audience reached (names, organizations, countries). Because of the many direct and indirect pathways through which fundamental research can impact applications, identifying the user audience and the research impacts can be very complex and time consuming.

**Objective:** The purpose of this paper is to describe a novel approach for identifying the pathways through which research can impact other research, technology development, and applications, and to identify the technical and infrastructure characteristics of the user population.

**Approach:** Citation Mining, a novel literature-based approach that integrates citation bibliometrics with text mining (extraction of useful information from text), was developed to identify the user community and its characteristics. Citation Mining starts with a group of core papers whose impact is to be examined, retrieves the papers that cite these core papers, and then analyzes the bibliometrics characteristics of the citing papers as well as their linguistic and thematic characteristics. The Science Citation Index is used as the source database for the core and citing papers, since its citation-based structure enables the capability to perform citation studies easily. The user community is characterized by the papers in the SCI that 1) cite the original research papers, and 2) cite the succeeding generations of these papers as well. Text mining is performed on the citing papers to identify the technical areas impacted by the research, the relationships among these technical areas, and relationships among the technical areas and the infrastructure (authors, journals, organizations). A key component of text mining, concept clustering, was used to provide both a taxonomy of the citing papers' technical themes and further technical insights based on theme relationships arising from the grouping process. Bibliometrics is performed on the citing papers to profile the user characteristics. In a specific example, Citation Mining is applied to the ~300 first generation citing papers of a fundamental physics paper on the dynamics of vibrating sand-piles.

**Results:** Most of the ~300 citing papers were basic research whose main themes were aligned with those of the cited paper. There were three main findings from a temporal analysis of the citing papers. First, the tail of total annual citation counts is very long, and shows little sign of abating. This is one characteristic feature of a seminal paper.

Second, the fraction of extra-discipline basic research citing papers to total citing papers ranges from about 15-25% annually, with no latency period evident. This lag-free extra-disciplinary diffusion may have been due to the combination of intrinsic broad-based applicability of the subject matter and publication of the paper in a high-circulation science journal with very broad-based readership. The text mining alone identified the intra-discipline applications and extra-discipline impacts and applications; this was confirmed by detailed reading of the ~300 abstracts.

Third, a four-year latency period exists prior to the emergence of the higher development category citing papers. This correlates with the results from the bibliometrics component. From the present study, it is not possible to differentiate the reasons for this important result. The latency could have been due to the inability of the technology community to *immediately* recognize the potential applications of the science. Or, it could have been due to the information remaining in the basic research journals, and not reaching the applications community. Or, the time that an application needs to be developed in this discipline is of the order of four years. Thus, the basic science publication feature that may have contributed heavily to extra-discipline citations may also have limited higher development category citations for the latency period.

**Conclusions:** The combination of citation bibliometrics and text mining provides a synergy unavailable with each approach taken independently. Furthermore, text mining is a REQUIREMENT for a feasible comprehensive research impact determination. The integrated multi-generation citation analysis required for broad research impact determination of highly cited papers will produce thousands or tens or hundreds of thousands of citing paper Abstracts. Text mining allows the impacts of research on advanced development categories and/ or extra-discipline categories to be obtained without having to read all these citing paper Abstracts. The multi-field bibliometrics provide multiple documented perspectives on the users of the research, and indicate whether the documented audience reached is the desired target audience.

## II. BACKGROUND

Identification of diverse research impacts is important to research managers, evaluators, and sponsors, and ultimately to performers. They are interested in the types of people and organizations citing the research outputs, and whether the citing audience is the target audience. Also, they are interested in whether the development categories and technical disciplines impacted by the research outputs are the desired targets. Since fundamental research can evolve along myriad paths, tracking diverse impacts becomes complex.

Presently, there are three generic approaches to tracking the impact of research: qualitative, semi-quantitative, and quantitative (Kostoff, 1997). Qualitative approaches are variants of peer review. Panels of experts are assembled, and impacts are identified based on the participants' knowledge, and usually personal experiences. The results are usually long on subjectivity, and short on independent documentation.

Semi-quantitative approaches are probably the most widely used for tracking impact (Kostoff, 1994). They include retrospective studies such as Hindsight (DOD, 1969) and Traces (IITRI, 1968), and various types of research sponsor accomplishment books such as those from DOE (DOE, 1983, 1986) and DARPA (IDA, 1991). A detailed treatment is contained in (Kostoff, 1997). Semi-quantitative approaches tend to be grounded in corporate memory of the participants, although some studies (Narin, 1989) follow the citation trail for supplementation. Their focus is detailed examination of a few high impact cases, rather than a wide-scale identification of many diverse impacts. As in the peer review approach, semi-quantitative approaches also have a high subjective component.

Quantitative approaches are also widely used for impact tracking (Kostoff, 1994, 1997). They tend to be divided between economic methods such as cost-benefit and internal rate-of-return (Averch, 1994; Tassey, 1999), and S&T indicators such as publications and patents (Narin, 1994), and their citations. They are the most objective of the three generic methods for tracking and quantifying research impact. However, many assumptions related to cost and benefit allocation are required for the economic studies (Kostoff, 1997). Additionally, many assumptions are required to accept correlation between numerical indicator values and degree of impact.

Thus, one of the gaps of all these impact tracking techniques is objective identification of the full scope of impacts produced by the research. These impacts include both the directly identifiable research impacts and the indirect impacts. For that fraction of performed research that is documented in the technical literature, tracking of direct and indirect research impacts on intermediate and final useful products becomes possible through tracking of generations of citations to the original research. If this wide scale impact information were obtained, then the in-depth studies performed by the semi-quantitative methods could cover an expanded range, or the roadmap of impacts could be presented as a self-contained valuable finding.

Even though the premier database for citation tracking, the Science Citation Index (SCI), contains a number of data fields abstracted from the full-text published papers, past citation-based studies using the SCI have focused almost exclusively on citation counts as an impact metric. Reviews of these citation studies can be found in (De Solla Price, 1986; Braun, 1987; Egghe, 1990). The potential impact of citation counts on decision-making is small, since the information content of citation counts alone is very limited. However, these citing records contain a wealth of information in their two main categories of diverse fields. The non-free-text fields, such as Author, Journal, Address, etc, describe the infrastructure characteristics of the citing community. The free-text fields, such as Title, Abstract, and Keywords (Keywords is not strictly a free-text field, but has sufficient technical characteristics to be included in this grouping), describe the technical characteristics of the impacted research, development, and applications areas.

Use of the SCI non-free-text fields for citing paper bibliometric analysis has been published on a very sporadic basis, and typically only for one or two data fields (Steele, 2000; Herring, 1999; Davidse, 1997). The focus of most of these studies has been on relating citations or citation rates to the few field variables examined. There do not appear to have been any citation studies performed for the specific purpose of user population profiling, where many of the available fields are examined in an integrated manner.

Recently, scientists have addressed the problem of citation in scientific research from a different perspective: looking for a topological description of citations (Bilke and Peterson, 2001), from power laws in citation networks (Redner, 1998), or power laws in number of cites received by journals according with their number of published papers (Katz, 2000) and finally trying to find some universal classes (Amaral et al. 2001). To overcome the limitations of these techniques, a phenomenological approach to deal with the information available and obtain a more detailed description of this complex system is presented in this paper.

Use of the SCI free-text fields for coupled trans-citation citing paper/ cited paper text mining analysis has not been published, although text mining studies of SCI and other database free-text fields have been reported (e.g., Kostoff et al, 2000a, 2000b, 2002, 2003).

## III. OBJECTIVES

The objectives of the present paper are:

i) Demonstrate the feasibility of tracking the myriad impacts of research on other research, development, and applications, using the technical literature.
ii) Demonstrate the feasibility of identifying a broad range of research product user characteristics, using the technical literature.
iii) Relate thematic characteristics of citing papers to their cited papers.

## IV. APPROACH

The present paper describes a novel process, Citation Mining (Kostoff et al, 2001a, Del Rio at al, 2002), that uses the best features of citation bibliometrics and text mining to track and document the impact of basic research on the larger R&D community across many generations. In Citation Mining, text mining (Kostoff et al, 2000a, 2000b, 2002, 2003; Losiewicz, 2000) of the cited and citing papers (trans-citation) supplements the information derived from the semi-structured field bibliometric analyses. Text mining illuminates the trans-citation thematic relationships, and provides insights of knowledge diffusion to other intra-discipline research, advanced intra-discipline development, and extra-discipline research and development. The addition of text mining to citation bibliometrics makes feasible the large-scale multi-generation citation studies that are necessary to display the full impacts of research.

A proof-of-principle demonstration of Citation Mining for user population profiling and research impact was performed on four sets of cited papers. The papers were selected based on the authors' technical interests, rather than a random representative sample. It was desired to have one group of papers representative of basic research, and another group representative of applied research. Two of the sets were selected Mexican and U. S. applied photo-voltaic research papers, and two of the sets were selected British and U. S. fundamental vibrating sand-pile research papers.

This paper presents the bibliometrics of those papers that cited all four sets of papers mentioned above, then focuses on the trans-citation coupled citing paper/ cited paper text mining results for one of the sets, a highly cited U. S. vibrating sand-pile paper (Jaeger, 1992). Vibrating sand-piles are important in their own right, since they model the behavior of granular systems used in agriculture (seeds, grains), geology (avalanches, soil mechanics), construction (gravel, sand), and manufacturing (powders, lubricants, sand-blasting). The underlying phenomena exhibited in their static and dynamic states can be found in many disparate applications, such as fusion confinement, geological formations, self-assembly of materials, thin film structure ordering, shock-wave statistics, and crowded airspace. Statistically, the sand-pile paper selected has sufficient citing papers for adequate text mining statistics. It covers an exciting area of physics research, and its technical sub-themes have potential for extrapolation to other technical disciplines.

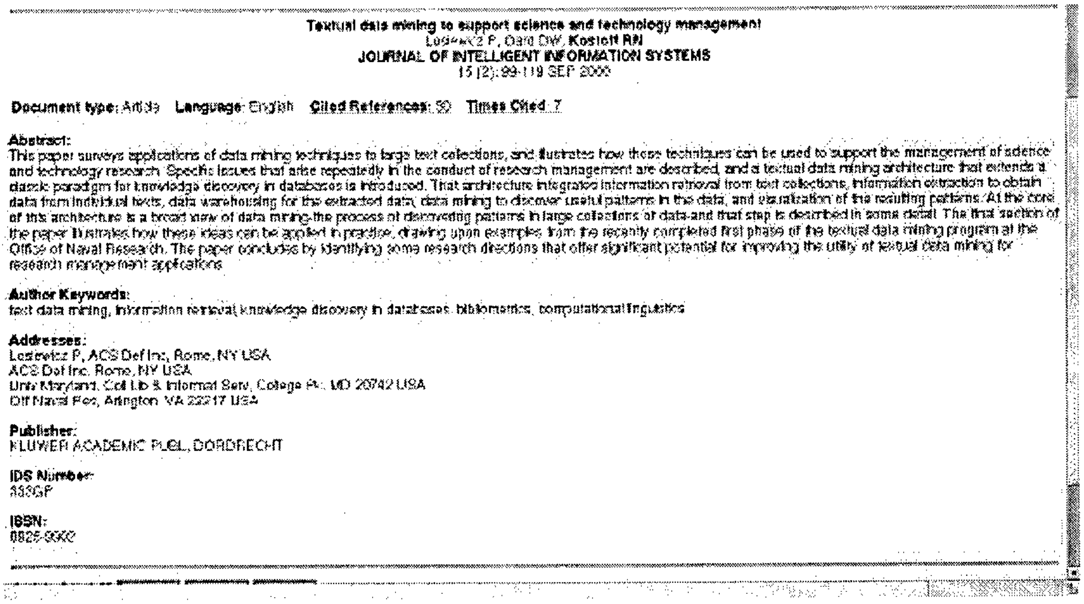
The analyses performed were of two types: bibliometrics and text mining. The text mining was subdivided into two components, manual concept clustering and statistical concept clustering. These different types of analyses are described in the following sections.

## IV-A. Bibliometrics Analysis

The citing paper summaries (records) were retrieved from the SCI. Analyses of the different non-free-text fields in each record were performed, to identify the infrastructure characteristics of the citing papers (authors, journals, institutions, countries, technical disciplines, etc).

This section starts by identifying the types of data contained in the SCI (circa early 2000), and the types of analyses that will be performed on this information (see Table 1).

FIGURE 1 – SAMPLE SCI RECORD



**Textual data mining to support science and technology management**
Losiewicz P, Oard DW, Kostoff RN
**JOURNAL OF INTELLIGENT INFORMATION SYSTEMS**
15 (2): 99-119 SEP 2000

**Document type:** Article **Language:** English **Cited References:** 50 **Times Cited:** 7

**Abstract:**
This paper surveys applications of data mining techniques to large text collections, and illustrates how those techniques can be used to support the management of science and technology research. Specific issues that arise repeatedly in the conduct of research management are described, and a textual data mining architecture that extends a classic paradigm for knowledge discovery in databases is introduced. That architecture integrates information retrieval from text collections, information extraction to obtain data from individual texts, data warehousing for the extracted data, data mining to discover useful patterns in the data, and visualization of the resulting patterns. At the core of this architecture is a broad view of data mining-the process of discovering patterns in large collections of data-and that step is described in some detail. The final section of the paper illustrates how these ideas can be applied in practice, drawing upon examples from the recently completed first phase of the textual data mining program at the Office of Naval Research. The paper concludes by identifying some research directions that offer significant potential for improving the utility of textual data mining for research management applications.

**Author Keywords:**
text data mining, information retrieval, knowledge discovery in databases, bibliometrics, computational linguistics

**Addresses:**
Losiewicz P, ACS Def Inc, Rome, NY USA
ACS Def Inc, Rome, NY USA
Univ Maryland, Coll Lib & Informat Serv, College Pk, MD 20742 USA
Off Naval Res, Arlington, VA 22217 USA

**Publisher:**
KLUWER ACADEMIC PUBL, DORDRECHT

**IDS Number:**
333GP

**ISSN:**
0925-9902

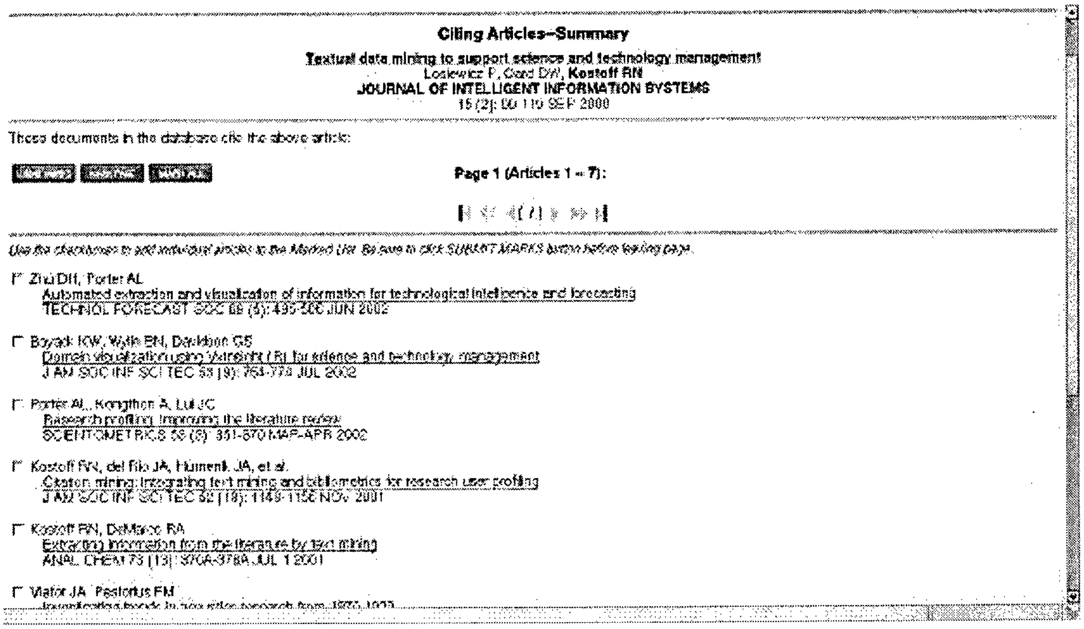
Figure 1 shows a sample record from the SCI. The actual paper that it represents is referred in the following description as the 'full paper'. Starting from the top, the individual fields are described in Table 1:

TABLE 1 – SCI RECORD FIELDS

1) Title - the complete title of the full paper.
2) Authors - all the authors of the full paper.
3) Source - journal name (e.g., Journal of Intelligent Information Systems).
4) Issue/ Page(s)/ Publication Date
5) Document Type - (e.g., Article, note, review, letter).
6) Language – the language of the full text document.

7) Cited References - the number and names of the references cited in the full paper
8) Times Cited - the number and names of the papers (whose records are contained in the SCI) that cited the full paper (see Figure 2). Thus, the number shown in this field is a lower bound.
9) Related Records – records that share one or more references (not shown).
10) Abstract - the complete Abstract from the full paper.
11) Author Keywords - keywords supplied by the author. In this example, no Keywords were supplied by the indexer, but the SCI contains a field for indexer Keywords, if supplied.
12) Addresses - organizational and street addresses of the authors. For multiple authors, this can be a difficult field to interpret accurately. Different authors from the same organizational unit may describe their organizational level differently. Different authors may abbreviate the same organizational unit differently.
13) Publisher

FIGURE 2. LIST OF CITING PAPERS OF ARTICLE SHOWN IN FIGURE 1.



**Citing Articles--Summary**

**Textual data mining to support science and technology management**
Losiewicz P, Oard DW, **Kostoff RN**
**JOURNAL OF INTELLIGENT INFORMATION SYSTEMS**
15 (2): 99-110 SEP 2000

These documents in the database cite the above article:

**Page 1 (Articles 1 -- 7):**

*Use the checkboxes to add individual articles to the Marked List. Be sure to click SUBMIT MARKS button before leaving page.*

□ Zhu DH, Porter AL
Automated extraction and visualization of information for technological intelligence and forecasting
TECHNOL FORECAST SOC 69 (5): 495-506 JUN 2002

□ Boyack KW, Wylie BN, Davidson GS
Domain visualization using VxInsight (R) for science and technology management
J AM SOC INF SCI TEC 53 (9): 764-774 JUL 2002

□ Porter AL, Kongthon A, Lui JC
Research profiling: Improving the literature review
SCIENTOMETRICS 53 (3): 351-370 MAR-APR 2002

□ Kostoff RN, del Rio JA, Humenik JA, et al.
Citation mining: Integrating text mining and bibliometrics for research user profiling
J AM SOC INF SCI TEC 52 (13): 1148-1156 NOV 2001

□ Kostoff RN, DeMarco RA
Extracting information from the literature by text mining
ANAL CHEM 73 (13): 370A-378A JUL 1 2001

□ Viator JA, Pestorius FM
Investigating trends in acoustics research from 1970-1999

How can the above fields be used in Citation Mining? In this paper, a phenomenological method to analyze the total information available in SCI database is proposed, as follows:

Title field is used in text mining together with the other unstructured text fields, Abstracts and Keywords, to perform the correlation analysis of the themes in the cited paper to those of the citing papers. Computational linguistics analysis is then performed.

Author field is used to obtain multi-author distribution profiles (e.g., number of papers with one author, number with two authors, etc).

Counts in Source field can lead to journal name distributions, theme distributions, and development level distributions.

Document Type register allows distributions of different document types to be computed (e.g., three articles, four conference proceedings, etc.).

Language field allows distributions over languages to be computed.

Cited References allows a historical analysis of the problem to be performed, and this field can be used to analyze the interrelations among different groups working on related problems.

Times Cited register would be important if the citing papers are of sufficient vintage. Then, their multiplier effect would be of interest, and could be computed. The distribution profile of times cited of the citing papers would be generated.

The Addresses register allows distributions of names and types of institutions, and countries, to be generated. Institution and country combinations would be of special interest, and could be correlated with author combination distributions.

The present demonstration of citation mining includes a comparison of a cited research unit from a developing country with a cited research unit from a developed country. It also compares a cited unit from a basic research field with a cited unit from an applied research field. Specifically, the technique is being demonstrated using selected papers from a Mexican semiconductor applied research group (MA), a United States semiconductor applied research group (UA), a British fundamental research group (BF), and a United States fundamental research group (UF) (see Table 2). These papers were selected based on the authors' familiarity with the topical matter, and the desire to examine papers that are reasonably cited. Sets of papers having at least 50 external cites were selected for analysis in order to have a good phenomenological description.

Table 2 – Cited Papers Used for Study

| GROUP | Times Cited | PAPERS |
|---|---|---|
| MA | 59 | Nair P.K. Sem. Sc. Tech. 3 (1988) 134-145<br>Nair P.K. J Phys D - Appl Phys, 22 (1989) 829-836<br>Nair M.T.S. Sem. Sc. and Tech. 4 (1989) 191- 199<br>Nair M. T. S. J Appl Phys, 75 (1994) 1557-1564 |
| UF | 307 | Jaeger HM, 1992, Science, V255, P1523 |
| BF | 119 | Mehta A, 1989, Physica A, V157, P1091<br>Mehta A, 1991, Phys Rev Lett, V67, P394<br>Barker GC, 1992, Phys Rev A, V45, P3435<br>Mehta A, 1996, Phys Rev E, V53, P92 |
| UA | 89 | Tuttle, Prog. Photovoltaic v3, 235 (1995)<br>Gabor, Appl. Phys. Lett. v65, 198 (1994)<br>Tuttle, J. Appl. Phys. v78, 269 (1995)<br>Tuttle, J. Appl. Phys. v77, 153 (1995)<br>Nelson, J. Appl. Phys. v74 5757 (1993) |

In addition, selection and banding of variables are key aspects of the bibliometric study. While specific variable values are of interest in some cases (e.g., names of specific citing institutions), there tends to be substantial value in meta-level groupings (e.g., institution class, such as government, industry, academia). Objectives of the study are to demonstrate important variables, types of meta-level groupings providing the most information and insight, and those conditions under which non-dimensionalization become useful. However, two analyses at the micro-level are presented involving specific correlations between both citing author and references for BF and UF papers. This latter analysis is directly important for the performers of scientific research. In addition, text mining could be performed on the text fields (mainly the Abstract, but including the Title and Keywords) to supplement the analysis on the semi-structured and structured fields (see Kostoff et al., 2000a, 2000b, 2001b, 2002, 2003).

## IV-B. Manual Concept Clustering

The purpose of the manual concept clustering was to generate a taxonomy (technical category classification scheme) of the database from the quantified technical phrases extracted from the free-text record fields. To generate the database, the citing papers' Abstracts were aggregated. Computational linguistics analyses were then performed on the aggregate. Technical phrases were extracted using the Database Tomography process (Kostoff et al, 1995, 2000a, 2000b; Losiewicz et al, 2000). An algorithm extracted all single, adjacent double, and adjacent triple word phrases from the text, and recorded the occurrence frequency of each phrase. While phrases containing trivial/ stop words at their beginning or end were eliminated by the algorithm, extensive manual processing was required to eliminate the low technical content phrases. Then, a taxonomy of technical sub-categories was generated by manually grouping these phrases into cohesive categories. Intra-discipline applications, and extra-discipline impacts and applications were identified from visual inspection of the phrases.

### IV-C. Statistical Concept Clustering

The purpose of the statistical concept clustering was to generate taxonomies of the database semi-automatically, again from the quantified technical phrases extracted from the free-text record fields. The clustering analysis further used quantified information about the relationships among the phrases from co-occurrence data (the number of times phrases occur together in some bounded domain). The statistical clustering analyses results complemented those from the manual concept clustering, and offered added perspectives on the thematic structure of the database.

After the phrase frequency analyses were completed, co-occurrence matrices of Abstract words and phrases (each matrix element Mij is the number of times phrase or word i occurs in the same record Abstract as phrase or word j) were generated using the TechOasis phrase extraction and matrix generation software. As in the phrase frequency analysis, the phrases extracted by the TechOasis natural language processor required detailed manual examination, to eliminate the low technical content phrases. The co-occurrence matrices were input to the WINSTAT statistical clustering software, where clusters (groups of related phrases based on co-occurrence frequencies) based on both single words and multi-word phrases were generated.

Two types of statistical clustering were performed, high and low level. The high level clustering used only the highest frequency technical phrases, and resulted in broad category descriptions. The low level clustering used low frequency phrases related to selected high frequency phrases, and resulted in more detailed descriptions of the contents of each broad category.

#### IV-C-1. High Level Clustering

The TechOasis phrase extraction from the citing Abstracts produced two types of lists. One list contained all single words (minus those filtered with a stop word list), and the other list contained similarly filtered phrases, both single and multi-word. Both lists required further manual clean-up, to insure that relatively high technical content material remained. The highest frequency items from each list were input separately to the TechOasis matrix generator, and two co-occurrence matrices, and resulting factor matrices, were generated.

The co-occurrence matrices were copied to an Excel file, and the matrix elements were non-dimensionalized. To generate clusters defining an overall taxonomy category structure for the citing papers, the Mutual Information Index was used as the dimensionless quantity. This indicator, the ratio of: the co-occurrence frequency between two phrases squared (Cij^2) to the product of the phrase occurrence frequencies (Ci*Cj), incorporates the co-occurrence of each phrase relative to its occurrence in the total text. The co-occurrence matrix row and column headings are arranged in order of decreasing frequency, with the highest frequency phrase occurring at the matrix origin. Based on the intrinsic nature of word and phrase frequencies, the row and column heading frequencies decrease rapidly with distance from the matrix origin. With increasing distance from the origin, the matrix becomes more and more sparse, although the phrases themselves have higher but more focused technical content. In parallel, the Mutual Information Index's values decrease rapidly as the distance from the matrix origin increases. Thus, the Mutual Information Index is useful for relating the highest frequency terms only, and for providing the top-level structural description of the taxonomy categories.

### IV-C-2. Low Level Clustering

To obtain a more detailed technical understanding of the clusters and their contents, the lower frequency phrases in each cluster need to be identified. A different matrix element non-dimensional quantity is required, one whose magnitudes remain relatively invariant to distance from the matrix origin. In addition, a different approach for clustering the low frequency phrases in the sparse matrix regions is required, one that relates the very detailed low frequency phrases to the more general high frequency phrases that define the cluster structure. In this way, the low frequency phrases can be placed in their appropriate cluster taxonomy categories.

The method chosen to identify the lower frequency phrases is as follows. Start with the cluster taxonomy structure defined by grouping the higher frequency phrases using the Average Neighbor agglomoration technique and the Mutual Information Index. Then, for each high frequency phrase in each cluster, find all phrases whose value of the Inclusion Index Ii exceeds some threshold. Ii is the ratio ofCij to Ci (the frequency of occurrence of phrase i in the total text), where phrase i has the lower frequency of the matrix element pair (i,j). A threshold value of 0.5 for Ii was used. The resultant lower frequency phrases identified by this method will occur rarely in the text, but when they do occur, they will be in close physical (and thematic) proximity to the higher frequency phrases.

## V. RESULTS

### V-A. Citation Bibliometrics

Figure 3
Authors Distribution Function

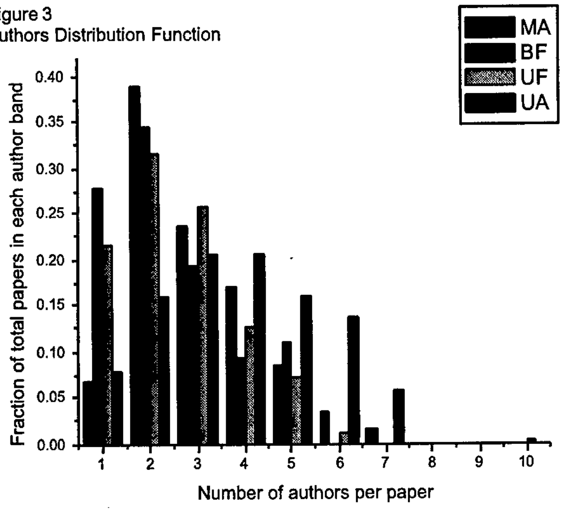
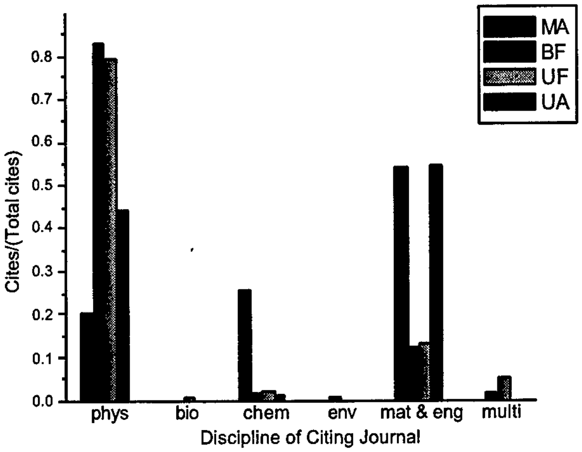


Figure 3 contains a bar graph of multi-author distribution for the four sets analyzed. The ordinate represents the fraction of total papers published in each author band, and the abscissa represents the number of authors per paper. The most striking feature of this graph is the behavior at the wings. The papers citing basic research dominate the low end (single author), while the papers citing applied research dominate the high end (6-7 authors). The papers citing basic research (BF and UF) have a similar number of authors per paper, with a maximum in the frequency distribution at two authors per paper. The UA citing papers show gaussian-like authorship distribution with three and four authors per paper, while the MA group citing papers show a distribution similar to the groups citing fundamental research papers but with fewer single-author papers. These four sets show author distributions where 90% of the papers had less than six authors. These results confirm the diversity of collaborative group compositions over different disciplines and levels of development.

Generally, as projects become more applied, they tend to become larger and more expensive, and require more resources. They also usually require the integration of multiple disciplines. Both these characteristics typically result in larger research groups, and hence in more contributors to a project and its resulting documents. Experimental work usually involves larger teams than theoretical work, while modeling and simulation activities tend to allow more individual efforts. The strong experimental emphasis of the two applied semiconductor groups, with little evidence of computer simulation shown, results in large teams on average. The more balanced theory/ experiment combination of the basic research group tends to suppress larger team efforts in favor of more individualized research. In addition, the intrinsic nature of sandpile vibration research, as opposed to elementary particle or fusion research, does not require large facilities and large research teams.

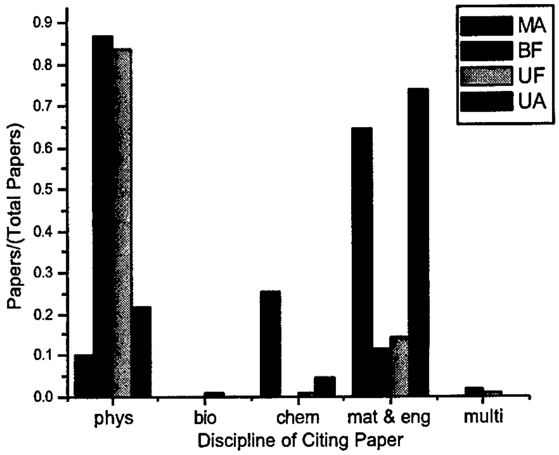
The citing journal discipline frequency is shown in Figure 4. Clearly, each paper set has defined its main discipline well. Also, there is a symmetry in the cross citing disciplines. UF and BF groups were cited more than 80% in fundamental journals and close to 10% in applied journals. Similarly, MA and UA groups were cited close to 50% in applied journals and 45% in fundamental journals. These journal discipline results suggest that the applications developed by the MA group have a strong impact on chemical journals, while the applications developed by the UA group strongly impact physics journals. A point to be stressed is that only the fundamental papers received cites in journals clearly outside of their disciplines.

Figure 4
Citing Journal Theme Distribution



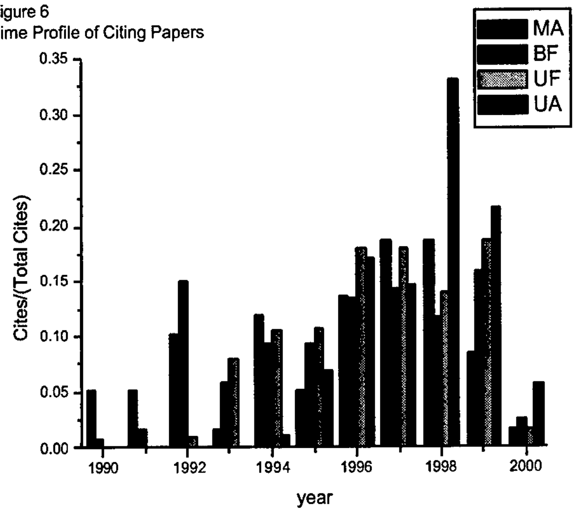
The discipline distribution of the citing papers, produced by analyzing the papers' Abstracts and Titles, is shown in Figure 5. It is slightly different from Figure 4. As concluded in the text mining, these free-text fields provide far more precise information than can be obtained from the journal discipline. Multi-disciplinary journals can publish uni-disciplinary papers from many different disciplines. Also, the journal categories, determined by ISI, are not a unique reflection of specific contents (e.g., an environmental journal can accept engineering papers, a materials journal can accept physics papers, etc). However, the chemical nature of the papers/ journals impacted by the MA group is confirmed.

Figure 5
Citing Paper Theme Distribution



In three of the four sets analyzed, the component papers were published in different years. The MA set was published from 1989 to 1994, UA from 1994 to 1995, BF from 1989 to 1996, while UF includes only one paper published in 1992. Figure 6 shows a clear oscillating behavior of UA and BF, due partly to the different dates of paper publication. Also, most of the sets have between 10% and 20% of cites per year, while the UA set received 38% ofthe cites in 1998.

Figure 6
Time Profile of Citing Papers



The single highly-cited paper feature of the UF set allows additional analyses and perspectives. In Figure 6a, the UF citing paper disciplines are shown as a function of time. As time evolves, citing papers from disciplines other than those of the cited paper emerge. An important point is the four-year delay of the systematic appearance of the more applied engineering and materials science citing papers.

Figure 6a
Time Profile of Citing Papers

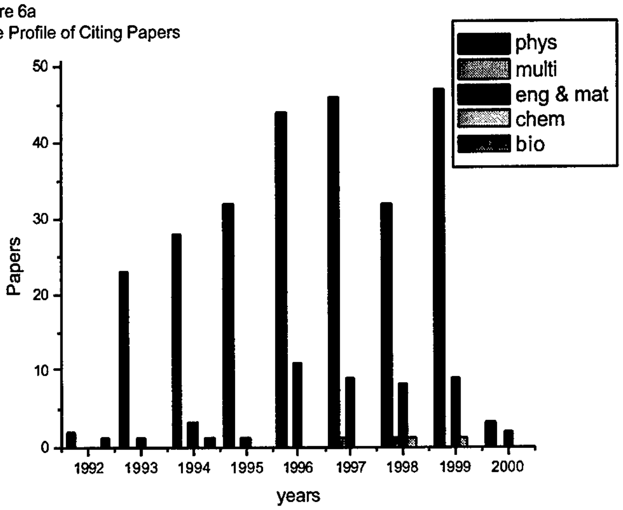


Figure 7 shows that most cites appear in articles. The four analyzed sets are cited in review articles and letters. This indicates the relevance of the analyzed papers. One important point is that only the fundamental papers are cited in notes, and only the UF paper was cited in an editorial document.

Figure 7
Citing Paper Type Profile

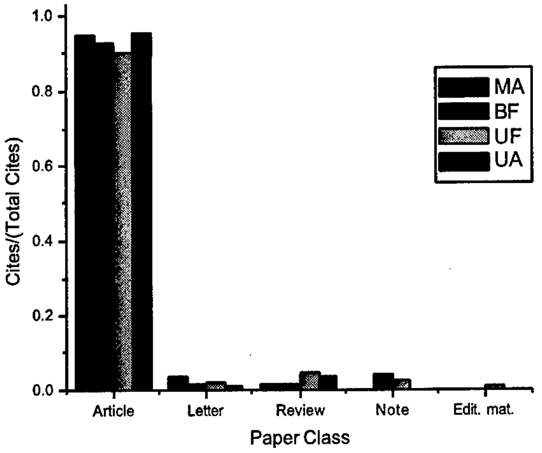


Figure 8 shows that English is the dominant language of all the paper sets analyzed. However, the surprising appearance of a significant number of citing papers written in Romanian for the MA set indicates that MA's work is important for at least one developing country. Also, there are no papers in Spanish.

Figure 8
Citing Paper Language Profile

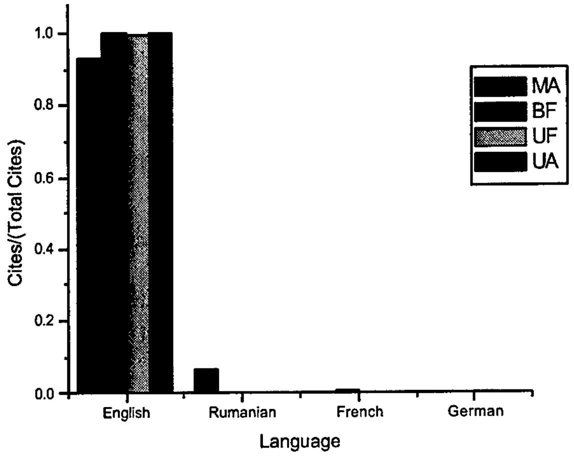
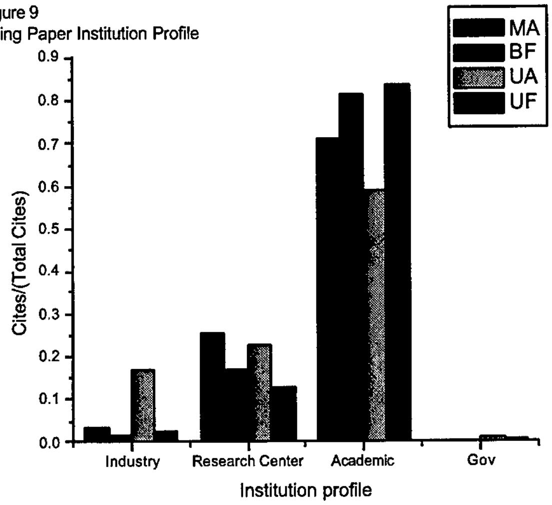


Figure 9 shows the profile of the citing institutions. Clearly, academia has the highest citing rates. Industry publications cite the advances in high-technological developments, but are not citing the advances in fundamental research. Research Centers follow applied and fundamental research about equally. Direct government participation is not significant in the fields studied. Government/ national laboratories were classified under research centers.

Figure 9
Citing Paper Institution Profile



There are 44 countries represented in the citing paper sets analyzed. Figure 10 shows only those countries with at least 10% of the citations for a set. USA has the most cites in aggregate. India has the largest cites of the MA set; Japan has the largest cites of the UA set. This fact is due to the different nature of the applied technology developed by MA and UA. The UA set contains work related to high technology, and the MA set is dedicated to explore low-cost technology. Therefore, this last set is cited by the less affluent countries of India, Romania and Mexico. India and Mexico also cite fundamental research, but not Romania. It is important to stress that if no low-cost technology papers were considered, these latter countries would not appear in this graph, and only developed countries would appear. Another point is that England does not cite UA works.

Figure 10
Citing Paper Country Profile

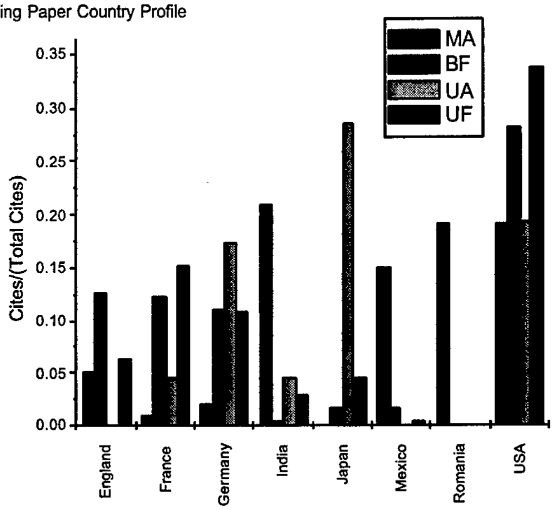
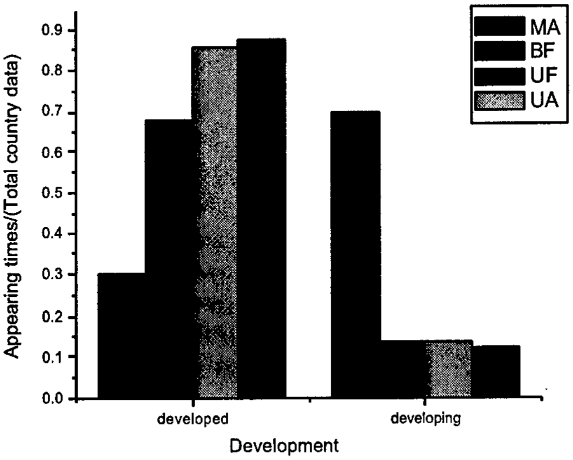


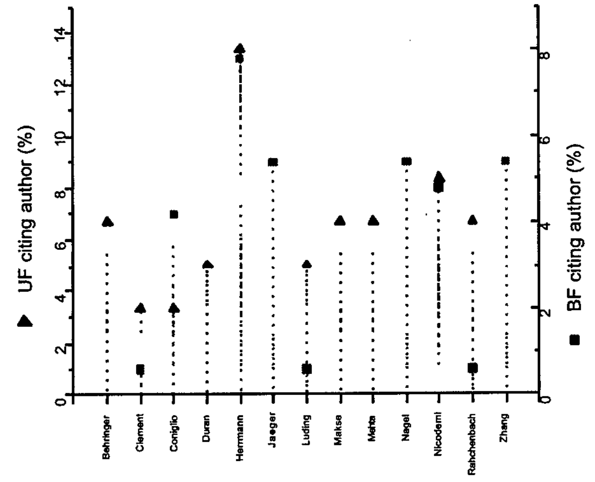
Figure 11 shows clearly that the low-cost technology papers are cited by developing countries. Developed countries cite the mostly high-technology papers.

Figure 11
Citing Country Development Phase Profile



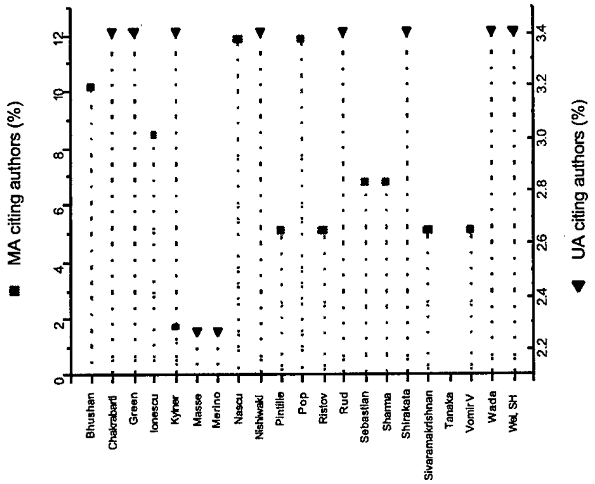
The analysis of the most common citing authors is presented in figures 12 and 13 where the frequency of an author citing UF (triangle) or BF (square) is plotted. Figure 12 shows that there is a close relation between the citing authors for both BF and UF groups. There is a common citing author who occupied the highest position in the frequency plot in both sets (Hermann, HJ). Three of the highest citing authors are not shared between the citing sets of UF and BF. Jaeger and Nagel are the authors of the UF paper and Mehta is one of the authors of BF paper. They maintain awareness of each other's work.

Figure 12



In contradistinction, Figure 13 shows that MA and UA have no intersection between their topics (low cost photovoltaic thin films and high efficient photovoltaic cells, respectively), from the perspective of the highest citing authors. Previous citation results have shown that applied research authors tend to cite more fundamental research, along relatively stratified lines. In Figure 13, it is clear that the maximum citing author of the MA group is a Romanian researcher.

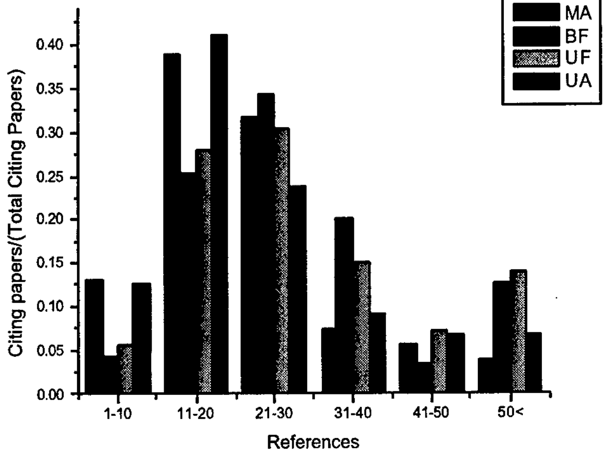
Figure 13



Tables A1 and A2 in the Appendix present the numerical data.

In Figure 14, it is clear that there are common features in the number of references in those papers that cite the core applied and fundamental papers, but there are also some differences. For instance, at the lower end of the spectrum (0-20), the applied papers' citing papers dominate. At the higher end of the spectrum (21-50+), the fundamental papers' citing papers dominate, with the exception of the BF anomaly at 41-50.

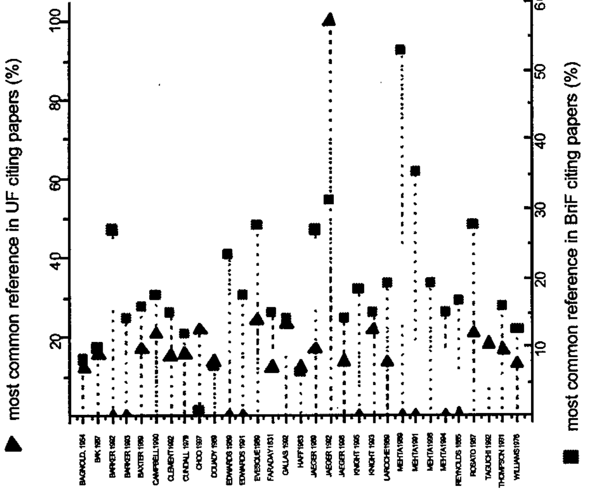
Figure 14
Citing Paper References Distribution



There are many possible reasons for these differences, and separating out the effects is complex. There are two different technical disciplines, and each one has its citing culture and traditions. Also, each technical discipline has a different level of research activity, and this could influence the magnitude of citations generated. Basic researchers tend to document more, and therefore produce a larger literature to cite. Finally, there may be different citing practices in basic and applied research.

Frequency analysis of the most common references in the citing papers provides insight to co-cited papers, and allows a historical perspective to be obtained. The reference-frequency for the UF and BF citing papers is shown in Figure 15. This figure shows clearly that the fundamental papers dealing with sand-piles are actually correlated.

Figure 15

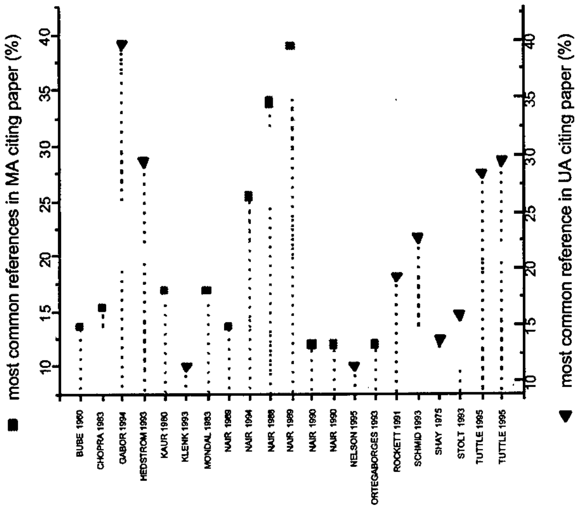


In this figure, Faraday's work (1831) appears within the twenty papers most cited in the UF and BF citing papers. This indicates the fundamental and seminal character of the experimental work performed by Faraday. Also, Reynolds' work (1885) appears within the twenty most cited papers in the references of the BF set. These two references also indicate the longevity of the unsolved problems tackled by the UF and BF groups.

The highest frequency co-cited papers have three interesting characteristics. They are essentially all in the same general physics area, they are all published in fundamental science journals (mainly physics), and they are all relatively recent, indicating a dynamic research area with high turnover. The detailed table is presented in the appendix.

The corresponding analysis of the most common references in the applied MA and UA groups is presented in figure 16. This figure shows clearly that these two groups have no correlations. However, in the detailed correlation analysis, there is one paper in the intersection of these two groups.

Figure 16



This ends the bibliometric analysis. The following section illustrates the usefulness of text mining analysis.

## V-B. Text Mining

### V-B-1. Manual Concept Clustering

The Abstract of the highly cited vibrating sand-pile paper (Jaeger, 1992) is shown in Figure 17.

FIGURE 17 – CITED PAPER ABSTRACT

Granular materials display a variety of behaviors that are in many ways different from those of other substances. They cannot be easily classified as either solids or liquids. This has prompted the generation of analogies between the physics found in a simple sandpile and that found in complicated microscopic systems, such as flux motion in superconductors or spin glasses. Recently, the unusual behavior of granular systems has led to a number of new theories and to a new era of experimentation on granular systems.

This paper had ~300 citing papers listed in the SCI, as of mid-CY2000. The highest frequency single, adjacent double, and adjacent triple word phrases from the aggregate citing papers (aligned with the central themes of the cited paper) can be represented by the following generic taxonomy: Theory/ modeling; Experiments/ measurements/ instruments/ variables; Structure/ Properties; Phenomena.

There were hundreds of technical phrases in each taxonomy category, and the authors selected those judged representative of each category for the purposes of illustration. Those representative phrases (*Underlined*) aligned with the central themes of the aggregate citing papers offer the following intra-discipline portrait of the citing aggregate. These papers reflect a balanced theoretical/ modeling effort (Molecular Dynamics Simulations, Monte Carlo Simulations, Kinetic Theory) and experimental effort (Magnetic Resonance Imaging, Charge Coupled Device Camera) targeted at studying the motions of granular particles. The papers focus on examination of the structure(s) and properties of vibrating sand-piles (Angle Of Repose, Coefficient Of Restitution), and the intrinsic phenomena of these collective systems (Collisions Between Particles, Fractional Brownian Motion), with emphasis on segregation (Size Segregation, Axial Segregation, Radial Segregation), relaxation (Relaxation Dynamics, Relaxation Time Tau), avalanching (Avalanch Durations, Avalanch Size), fluidization (Onset of Fluidization, Formation of Convection Cells), and collective behaviors (Collective Particle Motion, Self-Organized Criticality).

While the citing paper phrases mainly reflected emphasis on studies of granular piles, the phenomenological results and insights on segregation, relaxation, fluidization, avalanching, and collective behavior were extrapolated to some extra-discipline applications. These include (sample category abbreviated record Titles follow the phrases):

| Category | Phrases | Titles |
|---|---|---|
| geological formations and processes | (Earthquake*, Rock Avalanches, Carbonate Turbudite Deposition), | * *Sedimentary evolution of the early Paleocene deep-water Gulf of Biscay*<br>* *A fragmentation-spreading model for long-runout rock avalanches* |
| soil mechanics | (Soil Mechanics, Hillslope Gradient), | **Evidence for nonlinear, diffusive sediment transport on hillslopes and implications for landscape morphology*<br>**Analysis of vertical projectile penetration in granular soils* |
| industrial applications | (Screw Feeder*, Industrial), | **Precision dosing of powders by vibratory and screw feeders* |
| interacting object dynamics | (Traffic Congestion, War Game*). | **Study on crowded two-dimensional airspace - Self-organized criticality*<br>**Derivation and empirical validation of a* |

| | | |
|---|---|---|
| | | *refined traffic flow* |
| materials | (Rheolog*, Untwinned Single Crystals, Chemical Shift Tensors), | **Vortex avalanches at one thousandth the superconducting transition temperature*<br>**Mesoscale self-assembly of hexagonal plates using lateral capillary forces* |
| films | (Molecular Fluids, Adsorbed Polymer Layers), | **A model for the static friction behaviour of nanolubricated contacts*<br>**Spontaneous formation of ordered structures in thin films of metals* |
| multi-phase systems | (Flow Immunosensors, Fluidized Bed*), | **Advances in flow displacement immunoassay design*<br>**Rheophysical classification of concentrated suspensions and granular pastes*<br>**From bubbles to clusters in fluidized beds* |
| gas dynamics | (Gas Flow, Shock Waves, Shock Front), | **Statistics of shock waves in a two-dimensional granular flow*<br>**Scale invariant correlations in a driven dissipative gas* |
| micro particles | (Pollen Exine Sculpturing, Molecular, Spinule) | **The effects of genotype and ploidy level on pollen surface sculpturing in maize* |
| and microscale cooperative effects | (Tokamak, Plasma*, Lattice Gas). | **Sandpiles, silos and tokamak phenomenology: a brief review*<br>**Logarithmic relaxations in a random-field lattice gas subject to gravity* |

To validate the text mining results, each of the ~300 citing paper Abstracts was read by the first two authors, and those Abstracts reflecting applications and extra-discipline impacts were identified. *All of the applications and extra-discipline papers identified from reading the Abstracts could be identified/ retrieved from examination of the anomalous text mining-derived phrases with a threshold frequency of two.* The applications taxonomy of the previous section was validated using this Abstract reading and manual classification process, and judged to be a reasonable classification of the applications and extra-discipline impacts. Identification of the applications and extra-discipline impacts most unrelated to the main themes of the cited paper was easiest because of the highly anomalous nature of their representative phrases. Identification of the intra-discipline applications was the most difficult, since the phraseology used was similar to that of the cited paper themes.

***The importance of this result should be emphasized***. A complete citation impact study will typically involve multiple generations of citations. For a citation impact study that involves large numbers of first-generation citing papers and/ or large numbers of succeeding generation citing papers, reading each citing paper Abstract to identify applications paper characteristics becomes infeasible. For example, the ~300 citing papers of the sand-pile paper were themselves cited by ~3600 papers. If and when full-text becomes available for citation analysis, the time to read each paper will increase by an order of magnitude.

Use of text mining capabilities, such as computational linguistics, allows only those applications and extra-discipline papers of interest to be identified, and the requisite information could then be obtained from reading the Abstracts. In addition, the computational linguistics provides a structure and categorization of these myriad applications, allowing the larger context of application themes to be displayed and understood.

The citing papers representing categories of development and disciplines aligned and non-aligned with those of the cited paper are shown in the matrix of Figure 18.

FIGURE 18 – DEVELOPMENT CATEGORY AND CITED PAPER THEME ALIGNMENT OF CITING PAPERS

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| TECH DEV | 33 | | | | | | | | | |
| TECH DEV | 32 | | | | | 1 | | | | |
| TECH DEV | 31 | | | | | | | | | |
| APPL RES | 23 | | | | | | | 1 | 1 | 1 |
| APPL RES | 22 | | | | | 1 | | | 3 | |
| APPL RES | 21 | | | | | | 1 | 1 | | |
| BAS RES | 13 | 1 | 2 | 2 | 2 | 2 | 3 | | 1 | |
| BAS RES | 12 | | 2 | 3 | 6 | 4 | 10 | 8 | 10 | 1 |
| BAS RES | 11 | 3 | 23 | 28 | 27 | 43 | 43 | 30 | 33 | 4 |
| | | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 |

TIME

<u>CODE: MATRIX ELEMENT IS NUMBER OF PAPERS</u>

In Figure 18, the abscissa represents time. The ordinate, in the second column from the left, is a two-character tensor quantity. The first number represents the level of development characterized by the citing paper (1=basic research; 2=applied research; 3=advanced development/ applications), and the second number represents the degree of alignment between the main themes of the citing and cited papers (1=strong alignment; 2=partial alignment; 3=little alignment). Each matrix element represents the number of citing papers in each of the nine categories.

There are three interesting features on Figure 18. First, the tail of total annual citation counts is very long, and shows little sign of abating. This is one characteristic feature of a seminal paper.

Second, the fraction of extra-discipline basic research citing papers to total citing papers ranges from about 15-25% annually, with no latency period evident. This lag-free extra-disciplinary diffusion may have been due to the combination of intrinsic broad-based applicability of the subject matter and publication of the paper in a high-circulation science journal with very broad-based readership.

Third, a four-year latency period exists prior to the emergence of the higher development category citing papers. This correlates with the results from the bibliometrics component. From the present study, it is not possible to differentiate the reasons for this important result. The latency could have been due to the inability of the technology community to *immediately* recognize the potential applications of the science. Or, it could have been due to the information remaining in the basic research journals, and not reaching the applications community. Or, the time that an application needs to be developed in this discipline is of the order of four years. Thus, the basic science publication feature that may have contributed heavily to extra-discipline citations may also have limited higher development category citations for the latency period.

Here, only the development category characteristics of the first-generation directly citing papers have been considered. Conceivably, a higher development category paper could have cited one of the more fundamental first generation directly citing papers within the initial four-year latency period. These higher generation citing papers were beyond the scope of the present study.

Finally, the alignment of the citing journal theme to the main theme of the cited paper was estimated for all citing papers. In essentially all cases, the citing paper theme could be subsumed within the citing journal theme. However, given the breadth of most journal themes, this result had minimal information content (e.g., citing paper X was published in a Physics journal vs. a Materials journal).

In Davidse's study of Physics papers citations (Davidse, 1997), a key metric used in cross-disciplinary citations/ impacts was the distinction between Physics and non-Physics papers. It was implicitly assumed that the flow from Physics to non-Physics papers was analogous to the flow from basic to applied. While it may be true for some cases, Figure 18 (and other unpublished studies) shows that most extra-discipline flows in the present study were from basic physics research to basic research in the other disciplines. *Here, it is important to emphasize that a seminal idea gives new possible interpretations in many other disciplines.*

Davidse used journal themes (based on the SCI journal classification taxonomy) as a proxy for citing paper themes, with the level of resolution being at the gross discipline description, at best. There are many multi-discipline journals (e.g., Science, Nature, etc) that render a thematic distinction impossible. Davidse's approach required such a computerized proxy representation, since tens of thousands of citing papers were analyzed.

In contrast, the present paper's approach of identifying impact themes through text mining allows a much more detailed and informative picture of the impact of research to be obtained. It represents the difference between stating that a "Physics paper impacted Geology research" and a "paper focused on sand-pile avalanches for surface smoothing impacted analyses of steep hill-slope landslides".

## V-C. Statistical Concept Clustering

### V-C-1. High Level Clustering

For illustrative purposes, a sample truncated co-occurrence matrix based on phrases from the ~300 citing Abstracts is shown on Figure 19.

FIGURE 19 – PHRASE CO-OCCURRENCE MATRIX

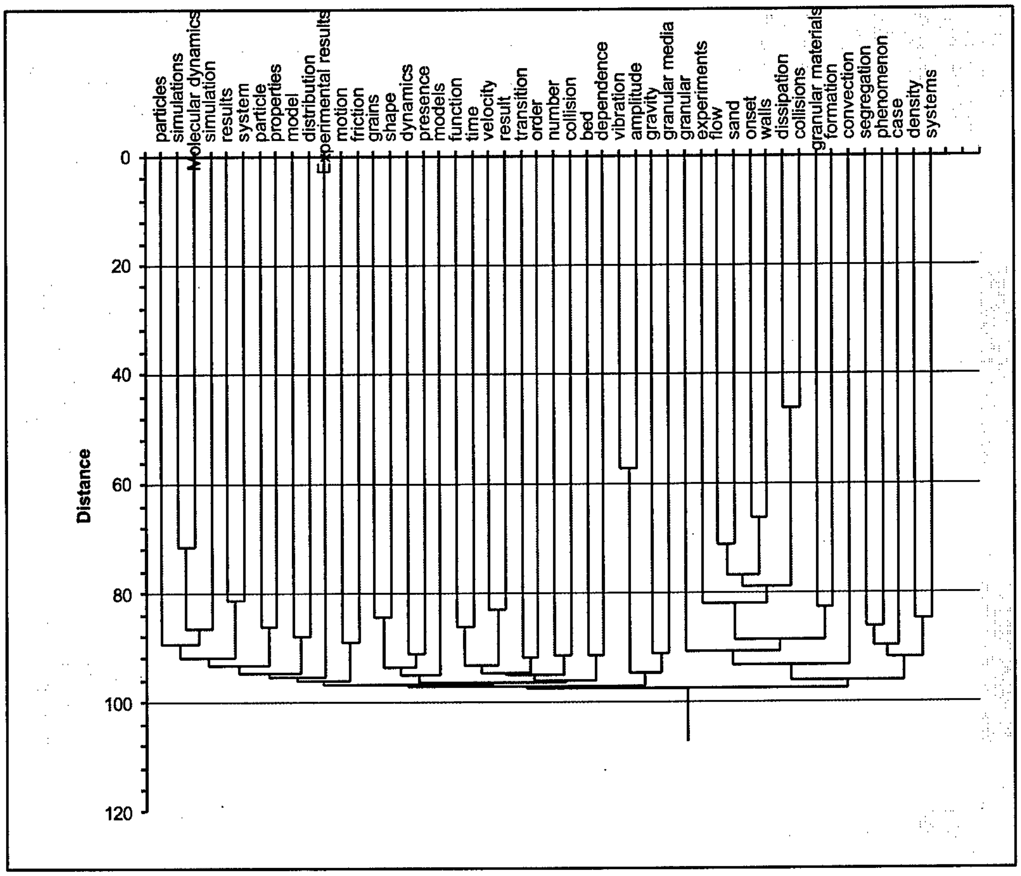
| # Records | 1 | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | particles | granular | results | system | Experiments | granular materials | grains | Flow | dynamics | motion | simulations | function | number | formation | segregation |
| 45 particles | 45 | 10 | 5 | 8 | 8 | 7 | 8 | 6 | 4 | 7 | 11 | 2 | 4 | 2 | 2 |
| 45 granular | 10 | 45 | 5 | 4 | 8 | 4 | 8 | 8 | 4 | 6 | 3 | | 4 | 5 | 1 |
| 40 results | 5 | 5 | 40 | 12 | 2 | 5 | 8 | 4 | 1 | 4 | 10 | 4 | 3 | 2 | 3 |
| 39 system | 8 | 4 | 12 | 39 | 4 | 1 | 9 | 2 | 9 | 6 | 10 | 3 | 6 | 3 | 2 |
| 37 experiments | 8 | 8 | 2 | 4 | 37 | 10 | 6 | 12 | 6 | 9 | 3 | 1 | 3 | 6 | 4 |
| 37 granular materials | 7 | 4 | 5 | 1 | 10 | 37 | 6 | 7 | 3 | 6 | 3 | 1 | 5 | 8 | 6 |
| 37 grains | 8 | 8 | 8 | 9 | 6 | 6 | 37 | 6 | 5 | 5 | 4 | 4 | 6 | 3 | 6 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 34 flow | 6 | 8 | 4 | 2 | 12 | 7 | 6 | 34 | 5 | 7 | 2 | 4 | 2 | 5 | 4 |
| 33 dynamics | 4 | 4 | 1 | 9 | 6 | 3 | 5 | 5 | 33 | 3 | 4 | 3 | 3 | 2 | 3 |
| 33 motion | 7 | 6 | 4 | 6 | 9 | 6 | 5 | 7 | 3 | 33 | 3 | 3 | 4 | 4 | 1 |
| 28 simulations | 11 | 3 | 10 | 10 | 3 | 3 | 4 | 2 | 4 | 3 | 28 | 2 | 2 | 1 | 4 |
| 25 function | 2 | | 4 | 3 | 1 | 1 | 4 | 4 | 3 | 3 | 2 | 25 | 3 | 1 | 2 |
| 21 number | 4 | 4 | 3 | 6 | 3 | 5 | 6 | 2 | 3 | 4 | 2 | 3 | 21 | 1 | 1 |
| 20 formation | 2 | 5 | 2 | 3 | 6 | 8 | 3 | 5 | 2 | 4 | 1 | 1 | 1 | 20 | 2 |
| 20 segregation | 2 | 1 | 3 | 2 | 4 | 6 | 6 | 4 | 3 | 1 | 4 | 2 | 1 | 2 | 20 |

In the final data analysis, a clustering of the 153 highest frequency technical content phrases in the matrix rows was then performed using the Excel add-in statistical package WINSTAT. A particularly helpful output for each clustering run was the dendogram, a tree-like diagram showing the structural branches that define the clusters. Figure 20 is one dendogram based on the 48 highest frequency phrases (for illustration purposes only). The abscissa contains the phrases that are clustered. The ordinate is a distance metric. The smaller the distance at which phrases, or phrase groups, are clustered, the closer is the connection between the phrases.

Thus, the first phrases combined are DISSIPATION and COLLISIONS, followed by VIBRATION and AMPLITUDE. At some later time, the VIBRATION-AMPLITUDE combination is grouped with the GRAVITY-GRANULAR MEDIA combination to form the next hierarchical level grouping, and so on. For the 48 phrases selected, the top hierarchical level consists of two clusters. On Figure 15, one cluster is bounded by the phrases PARTICLES-GRANULAR MEDIA, and the other is bounded by the phrases GRANULAR-SYSTEMS.

FIGURE 20 – PHRASE DENDOGRAM



Many agglomoration techniques were tested; the Average Neighbor method appeared to provide reasonably consistent good results. Analyses were performed of the numerous cluster options that were produced. The following is the top-level cluster description that represented the results of the phrase and word lists clustering best, as well as the factor matrix clustering from the TechOasis results.

The highest level categorization based on the highest frequency 153 phrases produced three distinct clusters: Structure/ Properties, Flow-Based Experiments, Modeling and Simulation. In the description of each cluster that follows, phrases that appeared within the clusters will be capitalized.

1) Structure/ Properties

This cluster contained MIXTURES of LARGE GRAINS and SMALL GRAINS, with STRATIFICATION along ALTERNATING LAYERS based on SIZE SEGREGATION and grain SHAPE and GEOMETRICAL PROFILE. The MIXTURE forms a PILE with an ANGLE of REPOSE. When the ANGLE of REPOSE is LARGER than a critical ANGLE, DYNAMICAL PROCESSES produce AVALANCHES, resulting in SURFACE FLOW within THIN LAYERS.

2) Flow-Based Experiments
This cluster contained EXPERIMENTS examining GRANULAR and SAND FLOW, The dependence of ENERGY DISSIPATION, due to COLLISIONS, on PACKING DENSITY was a focal area. The INFLUENCE of PIPE WALLS and PLATES on the SHEAR-driven VELOCITY and DENSITY PROFILES was studied, as well as ONSET of FLUIDIZATION and CONVECTIVE FLOW with its attendant FORMATION of CONVECTION CELLS.

3) Modeling and Simulation
This cluster contained MODELS and NUMERICAL SIMULATIONS based on EXPERIMENTAL RESULTS, OBSERVATIONS, MEASUREMENTS, and DATA. The SIMULATION METHODS MODELED the CHARACTERISTICS of DYNAMIC EVOLUTION from INITIAL CONDITIONS to STEADY STATE. A strong focal area was the CHARACTERISTICS of POWER SPECTRUM POWER LAW DISTRIBUTIONS, and their ROLE in the DYNAMIC EVOLUTION from INITIAL INSTABILITY to CRITICALITY. Sound PROPAGATION, especially its relation to DEPTH and PRESSURE, as a function of TIME and VIBRATION FREQUENCY, AMPLITUDE, and ACCELERATION is modeled with the statistical mechanics concepts of GRANULAR TEMPERATURE through KINETIC THEORY. Additionally, GRAVITY and VIBRATIONS are PHENOMENA used in the EQUATIONS to model the COMPACTION of GRANULAR MEDIA.

V-B-2. Low Level Clustering
Four types of results were obtained with the lower frequency phrases. Many of the lower frequency phrases were closely associated with one higher frequency phrase only; most lower frequency phrases were closely associated with one of the three clusters only; a few lower frequency phrases were associated with more than one cluster; and only a majority of the lower frequency phrases that related to applications or other disciplines were uniquely related to a single cluster. Sample relationships from each of these four types follow.

a) Lower Frequency Phrases Unique to One Higher Frequency Phrase
(High Frequency Phrase: Low Frequency Phrases)

REPOSE: VIBRATIONAL ACCELERATIONAL AMPLITUDE; STRATIFICATION: FACETED GRAINS; FLOW: VERTICAL GLASS PIPE, KINEMATIC SIEVING; COLLISIONS: LONG-RANGE CORRELATIONS; MODEL: COUPLED NONLINEAR STOCHASTIC EQUATIONS, SELF-ORGANIZED CRITICALITY; SIMULATION: DISCRETE ELEMENT METHOD; RELAXATION: STRONG SPATIAL CLUSTERING.

The phrases in this category, on average, tend to be longer and more detailed/ specific than the phrases in any of the other categories. They also tend to be the lowest frequency phrases, and their length and detail characteristics are consonant with the very lowest frequency phrases.

b) Lower Frequency Phrases Unique to One Cluster (Cluster High Frequency Phrases: Low Frequency Phrase)

LARGE GRAINS, SMALL GRAINS, REPOSE, STRATIFICATION: ALTERNATING LAYERS; COLLISIONS, CONVECTION CELLS, DISSIPATION EXPERIMENTS, FLOW, PACKING, VELOCITY PROFILES: ONSET OF FLUIDIZATION; DYNAMICS, RELAXATION: CONFIGURATIONAL ENTROPY; MODEL, SIMULATIONS: MKDV EQUATION

The low frequency phrases associated uniquely with the flow-based experiments cluster tended to be associated with the largest number of high frequency phrases, whereas the low frequency phrases associated uniquely with the modeling and simulation cluster tended to be associated with the smallest number of high frequency phrases. This reflects the more closely-knit nature of the flow-based experiments cluster relative to the more diverse nature of the modeling and simulation cluster, and was confirmed by examining all the high frequency phrases in each cluster.

c) Low Frequency Phrases Shared by All Three Clusters (High Frequency Phrases: Low Frequency Phrase)

POWER LAW, EXPERIMENTS, AVALANCHE: AVALANCHE DURATIONS; SIMULATIONS, EXPERIMENTS, STRATIFICATION: CONTACT NETWORK; DYNAMICS, ONSET, AVALANCHE: TOP LAYER; MODEL, FLOW, STRATIFICATION: STATIC GRAINS

As a general rule, the low frequency phrases in this category tend to be relatively generic, at least compared to phrases in the other three categories.

D) Low Frequency Phrases from Applications or other Disciplines (High Frequency Phrase(s): Low Frequency Phrase)
DENSITY WAVES: TRAFFIC FLOW; MODEL: AIR TRAFFIC; MODEL: CELL PELLETS; DYNAMICS, MODEL: DUNES; DYNAMICS, FLOW: IMMUNOSENSORS; MODEL, FLOW, AVALANCHES: GEOLOGICAL; MODEL, SIMULATION: WAR GAME; MODEL, DISSIPATION: VISCOELASTIC; GRANULAR TEMPERATURE: GAS FLUIDIZED BED; CONVECTION CELLS, EXPERIMENTS, FLOW, ONSET, VELOCITY PROFILES: TYPES OF RHEOLOGY

The clustering for relating themes and concepts is exceptionally complex. The categorization taxonomies, and subsequent allocations of phrases among the categories, are functions of the agglomoration technique, association metrics, phrase extraction algorithm, and interpretation of the results. In the present study, the highest level taxonomy was essentially invariant among these parameters, and was used for the examples. Interestingly, it was not substantially different from the highest level taxonomy obtained by visual inspection of the highest frequency phrases, as reported earlier in this paper. To obtain maximum benefits from what clustering can offer, lower categorical hierarchical levels must be accessed. More research is necessary to determine the most desirable combination of parameters to produce clusters at the lower hierarchical levels.

## VI. SUMMARY AND CONCLUSIONS

The first two objectives of this study were to demonstrate the feasibility of tracking the myriad impacts of research on other research, development, and applications, using the technical literature, and demonstrate the feasibility of identifying a broad range of research product user characteristics, using the technical literature. Both of these objectives were accomplished, along with some interesting technical insights about vibrating sandpile dynamics and temporal characteristics of information diffusion from research to applications. This wide range of results leads to the following conclusions.

Exploitation of the other types of information contained in the SCI and associated with the citation process offers the potential for providing R&D sponsors information that can help guide future directions of their R&D. In addition, the complete Citation Mining process described in the present paper has the potential to objectively document the breadth of impact of basic research on the R&D community. The addition of text mining to citation bibliometrics will make feasible the large-scale multi-generation citation studies that are necessary to display the full impacts of research.

Text mining is a requirement for making the total Citation Mining possible. Without text mining, either an overly general automated technique, such as journal classification, must be used to identify application areas, or tens or hundreds of thousands of Abstracts must be read. Text mining can locate small numbers of extra-discipline phrases (small signals) from large numbers of intra-discipline phrases (large clutter), and allow only those Abstracts of specific interest to be selected and read.

A substantial amount of human judgement and labor is required for all aspects of Citation Mining. For the bibliometric component of Citation Mining reported in this paper, classifying the results in groupings where judgement is required (e.g., Abstract technical theme, or applications theme) necessitates substantial work. For the text mining component described in detail in this paper, thousands of technical phrases must be examined. Judgements must be made as to their alignment with the main themes of the cited paper(s). Some of the bibliometric components conceivably could be automated (e.g., all the SCI journals could be classified by technical theme beforehand, then the alignment of the cited journal theme to the citing journal theme could be generated automatically). It is not clear how the selection of extra-discipline phrases could be automated, given the intense expert judgement required.

The third of the study objectives was to relate thematic characteristics of citing papers to their cited papers. There was a strong relation of these thematic characteristics for the sandpile paper and its citing papers, and an even stronger citing/ cited paper relationship for the applied photo-voltaic research papers. This result has potential far-reaching implications for the corporate and national security intelligence communities. Through the tracking of cited papers, one could theoretically infer the theme(s) of the citing papers, and vice versa. Very little has been reported in the literature on this broader field of trans-citation analysis, especially using text mining as reported in the present paper, and the broader field is ripe for further research and exploitation.

This study referred to, but did not examine details of, second or higher generation citations. The authors believe they are valid measures or indicators of influence and impact, but the actual method of impact quantification remains an open question. More research is required to understand the principles of allocating impact among a paper's references.

Finally, there is a very important message that emerges from the results of the present study relative to the sponsorship of basic research. Over the past decade, the trend in industry and government has been toward requirements-driven research (e.g., the term 'strategic research' is becoming used more widely in government agencies, and corporately-funded industrial research has strongly evolved into profit-center sponsored research). While this may be beneficial to the sponsoring organization from a short-term tactical perspective, the long-term strategic perspective may suffer. Would fundamental sand-pile research receive funding from Tokamak, air traffic control, or materials programs, even though sand-pile research could impact these or many other types of applications, as shown in this paper? *It is necessary to stress that sponsorship of some unfettered research must be protected, for the strategic long-term benefits on global technology and applications!*

Acknowledgement
This paper has been partially supported by DGAPA-UNAM under Project IN103100.

## VII. REFERENCES

Amaral, LAN, Gopikrishnan, P., Matia, K., Plerou, V. and Stanley E.H. (2001). Application of statistical physics methods and concepts to the study of science & technology systems, Scientometrics, 51, 3

Averch, H., (1994) "Economic Approaches to the Evaluation of Research", Evaluation Review, 18:1, February, p. 77-88.

Bilke, S. and Peterson, C. (2001). Topological properties of citation and metabolic networks, Phys. Rev. E 64, 036106

Braun, T. et al, (1987), "Literature of Analytical Chemistry. A Scientometric Evaluation", CRC Press.

Davidse, R. J., and Van Raan, A. F. J., (1997). "Out of Particles: Impact of CERN, DESY, and SLAC Research to Fields other than Physics", Scientometrics, 40:2. P. 171-193.

De Solla Price, D. J., (1986) "Little Science, Big Science and Beyond", Columbia University Press.

Del Río, J. A., Kostoff, R. N., García, E. O., Ramírez, A. M., and Humenik, J. A., (2000). "Citation Mining: Citing Population Profiling using Bibliometrics and Text

Mining". Centro de Investigación en Energía, Universidad Nacional Autonoma de Mexico http://www.cie.unam.mx/xml/tc/ft/arp/citing.xml.

Del Rio, J. A., Kostoff, R. N., Garcia, E. O., Ramirez, A. M., and Humenik, J. A. (2002). "Phenomenological Approach to Profile Impact of Scientific Research: Citation Mining." Advances in Complex Systems. 5:1. 19-42.

DOD, (1969) Project Hindsight, Office of the Director of Defense Research and Engineering, Wash., D. C., DTIC No. AD495905, October.

DOE, (1983) "Health and Environmental Research: Summary of Accomplishments", Office of Energy Research, Office of Program Analysis, Report No. DOE/ER-0194, May; DOE, (1986) "Health and Environmental Research: Summary of Accomplishments", Office of Energy Research, Office of Program Analysis, Report No. DOE/ER-0275, August.

Egghe, L., and Rousseau, R., (1990) "Introduction to Informetrics", Elsevier.

Herring, S. D., (1999). "The Value of Inter-disciplinarity: A Study Based on the Design of Internet Search Engines", JASIS, 1 April, p. 358-365.

IDA, (1991) "DARPA Technical Accomplishments", Volume I, IDA Paper P-2192, February 1990; Volume II, IDA Paper P-2429, April 1991; Volume III, IDA Paper P-2538, July 1991, Institute for Defense Analysis.

IITRI, (1968) "Technology in Retrospect and Critical Events in Science", Illinois Institute of Technology Research Institute Report, December.

Jaeger, H. M., and Nagel, S. R. (1992). "Physics of the Granular State". Science. 256. 20 March, p. 1523-1531.

Katz, J.S. (2000) "Scale-independent Indicators and Research Evaluation" Science and Public Policy. 27. I. 23-36.

Kostoff, R. N., (1994) "Assessing Research Impact: US Government Retrospective and Quantitative Approaches", Science and Public Policy, 21:1, February.

Kostoff, R. N., Eberhart, H. J., and Miles, D. (1995). "System and Method for Database Tomography", U. S. Patent Number 5440481, August 8.

Kostoff, R. N., (1997) "The Handbook of Research Impact Assessment", DTIC Report Number ADA296021, Summer. Also, see http://www.dtic.mil/dtic/kostoff/index.html.

Kostoff, R. N., Braun, T., Schubert, A., Toothman, D. R., and Humenik, J. (2000a). "Fullerene Roadmaps Using Bibliometrics and Database Tomography". Journal of Chemical Information and Computer Science. 40:1. Jan-Feb. p. 19-39.

Kostoff, R. N., Green, K. A., Toothman, D. R. Humenik, J. A., (2000b) "Database Tomography Applied to an Aircraft Science and Technology investment Strategy", Journal of Aircraft, 37:4, p. 727-730.

Kostoff, R. N., Del Rio, J. A., García, E. O., Ramírez, A. M., and Humenik, J. A. (2001a). "Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling". JASIST. 52:13. 1148-1156. 52:13. November.

Kostoff, R. N., and DeMarco, R. A. (2001b). "Science and Technology Text Mining". Analytical Chemistry. 73:13. 370-378A. 1 July.

Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. (2002). "Electrochemical Power Source Roadmaps using Bibliometrics and Database Tomography". Journal of Power Sources. 110:1. 163-176.

Kostoff, R. N., Shlesinger, M., and Malpohl, G. (2003). "Fractals Roadmaps using Bibliometrics and Database Tomography". Fractals. December.

Losiewicz, P., Oard, D., and Kostoff, R. N. (2000). "Textual Data Mining to Support Science and Technology Management". Journal of Intelligent Information Systems. 15:2, p. 99-119.

Narin, F., (1989) "The Impact of Different Modes of Research Funding", in:Evered, David and Harnett, Sara, Eds., The Evaluation of Scientific Research, John Wiley and Sons, Chichester, UK, p. 120-140.

Narin, F., Olivastro, D., and Stevens, K. A., (1994) 'Bibliometrics -Theory, Practice, and Problems", Evaluation Review, 18:1, February, p. 65-76.

Redner, S. (1998). "How Popular is your Paper? An Emperical Study of the Citation Distribution." Eur. Phys. J. B 4, 131.

Steele, T. W., (2000). “The Impact of Interdisciplinary Research in the Environmental Science: A Forestry Case Study”, JASIS, 15 March, p. 476-484.

Tassey, G., (1999) “Lessons Learned About The Methodology Of Economic Impact Studies: The NIST Experience,” Evaluation and Program Planning, **22, p.** 113—119.

## **VIII.** APPENDIX

Tables A1 to A4 contain the most frequent citing authors for the four sets of papers.

### TABLE A1 – BF CITING AUTHORS

| BF Citing Authors | | |
|---|---|---|
| Citing Author | Citing Times | Percentage |
| Herrmann, HJ | 16 | 13 |
| Jaeger, HM | 11 | 9 |
| Nagel, SR | 11 | 9 |
| Zhang, ZP | 11 | 9 |
| Nicodemi, M | 10 | 8 |

### TABLE A2 – UF CITING AUTHORS

| UF Citing Authors | | |
|---|---|---|
| Citing Author | Citing Times | Percentage |
| Herrmann, HJ | 24 | 8 |
| Nicodemi, M | 14 | 5 |
| Rahchenbach, J | 11 | 4 |
| Mehta, A | 11 | 4 |
| Makse, HA | 11 | 4 |
| Behringer, RP | 11 | 4 |
| Duran, J | 10 | 3 |
| Luding, S | 9 | 3 |
| Coniglio, A | 8 | 2 |

| | | |
|---|---|---|
| Clement, E | 8 | 2 |

## TABLE A3 – MA CITING AUTHORS

| MA Citing Authors | | |
|---|---|---|
| Citing Author | Citing Times | Percentage |
| Nascu C | 7 | 0.12 |
| Pop I | 7 | 0.12 |
| Bhushan S | 6 | 0.10 |
| Ionescu V | 5 | 0.08 |

## TABLE A4 UA Citing Authors

| UA Citing Authors | | |
|---|---|---|
| Citing Author | Citing Times | Percentage |
| Rud, VY | 8 | 9 |
| Wada, T | 8 | 9 |
| Negami, T | 7 | 8 |
| ZUNGER, A | 6 | 7 |
| Kohara, N | 5 | 6 |
| Schock, HW | 5 | 6 |
| Tanaka, T | 5 | 6 |
| Yamaguchi, T | 5 | 6 |
| Yoshida, A | 5 | 6 |

Tables A5 to A8 contain frequencies of most cited papers in the citing papers of the four different sets.

### TABLE A5 – FREQUENCIES OF REFERENCES IN BF CITING PAPERS

| Frequencies of References in BF Citing Papers | | |
|---|---|---|
| Paper | Times | |
| MEHTA A, 1989, PHYSICA A, V157, P1091 | 63 | 52.9% |
| MEHTA A, 1991, PHYS REV LETT, V67, P394 | 42 | 35.3% |
| JAEGER HM, 1992, SCIENCE, V255, P1523 | 37 | 31.1% |

| | | |
|---|---|---|
| EVESQUE P, 1989, PHYS REV LETT, V62, P44 | 33 | 27.7% |
| ROSATO A, 1987, PHYS REV LETT, V58, P1038 | 33 | 27.7% |
| BARKER GC, 1992, PHYS REV A, V45, P3435 | 32 | 26.9% |
| JAEGER HM, 1989, PHYS REV LETT, V62, P40 | 32 | 26.9% |
| EDWARDS SF, 1989, PHYSICA A, V157, P1080 | 28 | 23.5% |
| LAROCHE C, 1989, J PHYS-PARIS, V50, P699 | 23 | 19.3% |
| MEHTA A, 1996, PHYS REV E, V53, P92 | 23 | 19.3% |
| KNIGHT JB, 1995, PHYS REV E, V51, P3957 | 22 | 18.5% |
| CAMPBELL CS, 1990, ANNU REV FLUID MECH, V22, P57 | 21 | 17.6% |
| EDWARDS SF, 1991, J STAT PHYS, V62, P889 | 21 | 17.6% |
| REYNOLDS O, 1885, PHILOS MAG 5, V20, P469 | 20 | 16.8% |
| BAXTER GW, 1989, PHYS REV LETT, V62, P2825 | 19 | 16.0% |
| THOMPSON PA, 1991, PHYS REV LETT, V67, P1751 | 19 | 16.0% |
| CLEMENT E, 1991, EUROPHYS LETT, V16, P133 | 18 | 15.1% |
| FARADAY M, 1831, PHIL T R SOC LONDON, V52, P299 | 18 | 15.1% |
| KNIGHT JB, 1993, PHYS REV LETT, V70, P3728 | 18 | 15.1% |
| MEHTA A, 1994, GRANULAR MATTER | 18 | 15.1% |
| BARKER GC, 1993, PHYS REV E, V47, P184 | 17 | 14.3% |
| GALLAS JAC, 1992, PHYS REV LETT, V69, P1371 | 17 | 14.3% |
| JAEGER HM, 1996, REV MOD PHYS, V68, P1259 | 17 | 14.3% |

### TABLE A6 – FREQUENCIES OF REFERENCES IN UA CITING PAPERS

| Frequencies of References in UA Citing Papers | | |
|---|---|---|
| Paper | Times | |
| GABOR AM, 1994, APPL PHYS LETT, V65, P198 | 35 | 39.8% |
| HEDSTROM J, 1993, P 23 IEEE PHOT SPEC, P364 | 26 | 29.5% |
| TUTTLE JR, 1995, PROG PHOTOVOLTAICS, V3, P383 | 26 | 29.5% |
| TUTTLE JR, 1995, J APPL PHYS, V77, P153 | 25 | 28.4% |
| SCHMID D, 1993, J APPL PHYS, V73, P2902 | 20 | 22.7% |
| ROCKETT A, 1991, J APPL PHYS, V70, PR81 | 17 | 19.3% |
| STOLT L, 1993, APPL PHYS LETT, V62, P597 | 14 | 15.9% |
| SHAY JL, 1975, TERNARY CHALCOPYRITE | 12 | 13.6% |
| KLENK R, 1993, ADV MATER, V5, P144 | 10 | 11.4% |
| NELSON AJ, 1995, J APPL PHYS, V78, P269 | 10 | 11.4% |
| BOEHNKE UC, 1987, J MATER SCI, V22, P1635 | 9 | 10.2% |
| CONTRERAS MA, 1994, PROG PHOTOVOLTAICS R, V2, P287 | 9 | 10.2% |
| FEARHEILEY ML, 1986, SOL CELLS, V16, P91 | 9 | 10.2% |
| JAFFE JE, 1984, PHYS REV B, V29, P1882 | 8 | 9.1% |
| NELSON AJ, 1993, J APPL PHYS, V74, P5757 | 8 | 9.1% |
| TUTTLE JR, 1996, MATER RES SOC SYMP P, V426, P143 | 8 | 9.1% |

### TABLE A7 – FREQUENCIES OF REFERENCES IN MA CITING PAPERS

| Frequencies of References in MA Citing Papers | | |
|---|---|---|
| Paper | Times | |
| NAIR PK, 1989, J PHYS D APPL PHYS, V22, P829 | 23 | 25.84% |
| NAIR PK, 1988, SEMICOND SCI TECH, V3, P134 | 20 | 22.47% |
| NAIR MTS, 1994, J APPL PHYS, V75, P1557 | 15 | 16.85% |
| KAUR I, 1980, J ELECTROCHEM SOC, V127, P943 | 10 | 11.24% |
| MONDAL A, 1983, SOL ENERG MATER, V7, P431 | 10 | 11.24% |
| CHOPRA KL, 1983, THIN FILM SOLAR CELL | 9 | 10.11% |

| | | |
|---|---|---|
| BUBE RH, 1960, PHOTOCONDUCTIVITY SO | 8 | 8.99% |
| NAIR MTS, 1989, SEMICOND SCI TECH, V4, P191 | 8 | 8.99% |

## TABLE A8 – FREQUENCIES OF REFERENCES IN UF CITING PAPERS

| Frequencies of References in UF Citing Papers | | |
|---|---|---|
| Paper | Times | |
| JAEGER HM, 1992, SCIENCE, V255, P1523 | 307 | 100% |
| EVESQUE P, 1989, PHYS REV LETT, V62, P44 | 75 | 24.4% |
| GALLAS JAC, 1992, PHYS REV LETT, V69, P1371 | 72 | 23.4% |
| CHOO K, 1997, PHYS REV LETT, V79, P2975 | 68 | 22.1% |
| KNIGHT JB, 1993, PHYS REV LETT, V70, P3728 | 68 | 22.1% |
| ROSATO A, 1987, PHYS REV LETT, V58, P1038 | 64 | 20.8% |
| CAMPBELL CS, 1990, ANNU REV FLUID MECH, V22, P57 | 62 | 20.8% |
| TAGUCHI YH, 1992, PHYS REV LETT, V69, P1367 | 56 | 18.2% |
| JAEGER HM, 1989, PHYS REV LETT, V62, P40 | 52 | 16.9% |
| BAXTER GW, 1989, PHYS REV LETT, V62, P2825 | 52 | 16.9% |
| THOMPSON PA, 1991, PHYS REV LETT, V67, P1751 | 51 | 16.6% |
| BAK P, 1987, PHYS REV LETT, V59, P381 | 48 | 15.6% |
| CUNDALL PA, 1979, GEOTECHNIQUE, V29, P47 | 48 | 15.6% |
| CLEMENT E, 1992, PHYS REV LETT, V69, P1189 | 47 | 15.3% |
| JAEGER HM, 1996, REV MOD PHYS, V68, P1259 | 43 | 14.0% |
| DOUADY S, 1989, EUROPHYS LETT, V8, P621 | 43 | 14.0% |
| LAROCHE C, 1989, J PHYS-PARIS, V50, P669 | 42 | 13.7% |
| WILLIAMS JC, 1976, POWDER TECHNOL, V15, P 245 | 41 | 13.4% |
| HAFF PK, 1983, J FLUID MECH, V134, P401 | 38 | 12.4% |
| FARADAY M, 1831, PHIL T R SOC LONDON, V52, P299 | 37 | 12.5% |
| BAGNOLD RA, 1954, P ROY SOC LOND A MAT, V225, P49 | 37 | 12.5% |